

A method and apparatus are disclosed for compressing textual documents encoded using a tag-based markup language,, such as XML or SGML documents, in a manner that allows a compressed document to be processed without decompression. A document is compressed using a standard compression algorithm that is applied only to the data elements of the document. The tag elements of the XML document that impose the hierarchical structure on the data elements are not compressed. The hierarchical relationship of the data elements can be ascertained from the compressed document. A user can thereafter selectively decompress desired portions of the document, without decompressing the entire document. An identification of the employed compression technique can be inserted into a root node tag element of the XML document or into an additional tag element pair that is inserted into the XML document. An XML decoder can utilize the uncompressed tag elements to ascertain the hierarchical structure of the compressed data and present the user with a corresponding hierarchical expression of the document.

2020-11-25 15:45